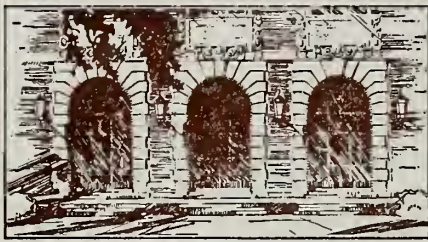


LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

510.84
I l 6 r
no. 679-684
cop. 2



The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

NOV 5 1977

OCT 13 REC'D

JAN 23 1981

510.84

Ilbr

UIUCDCS-R-74-679

u.679

DEC 4 1974

ON RANDOM 3-2 TREES

by

Andrew Chi-Chih Yao

October 1974



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

THE LIBRARY OF THE

NOV 27 1974

UNIVERSITY OF ILLINOIS



151
UIUCDCS-R-74-679

On Random 3-2 Trees

by

Andrew Chi-Chih Yao

October 1974

Department of Computer Science
University of Illinois
Urbana, Illinois 61801

Research supported by NSF Grant GJ41538.



Digitized by the Internet Archive
in 2013

<http://archive.org/details/onrandom32trees679yaoa>

510.84
Ilbr
no 679-684
cgr 2

ABSTRACT

It is shown that $\bar{n}(N)$, the average number of nodes in an N -key random 3-2 tree, satisfies the inequality $0.70N - \epsilon < \bar{n}(N) < 0.79N + \epsilon$. A similar analysis is done for general B-trees. It is shown that storage utilization is essentially $\ln 2 \approx 69\%$ for B-trees of high orders.

1. Introduction

Balanced tree structures are often used in the organization of information. One attractive scheme, called "3-2 trees," was introduced by J. Hopcroft [1] [4]. Some interesting questions concerning 3-2 trees have been raised [3] [4]. In this paper we present a partial solution to a problem posed in [3].

A 3-2 tree is a tree in which every internal node contains either 1 or 2 keys, and all the leaves are at the same level (see Figure 1). In drawing 3-2 trees we shall adopt the notation used in [3]. Thus keys are represented by dots inside a node as we shall only be interested in the structure of the trees.

To put a new key into a node that contains only one key, we simply insert it as the second key. If the node already contains 2 keys, we split the node into two nodes containing respectively the minimum and the maximum of the three keys, and insert the middle key into the parent node by repeating the process. When there is no node above, a new root node will be created to hold the middle key.

Consider a 3-2 tree T with $j-1$ keys in it. These $j-1$ keys divide all possible key values into j intervals. The insertion of a new key K_j into T is said to be a random insertion if K_j has equal probabilities for being in any one of the j intervals defined above.

Now consider the building of 3-2 trees by successive random insertions. The average cost involved is dependent on the specific implementation of the insertion algorithm. There are, however, certain quantities that are useful in general for the analysis. One quantity of interest is $\bar{n}(N)$, the average number of internal nodes in a 3-2 tree after N keys have been randomly inserted into the empty tree. In this paper we

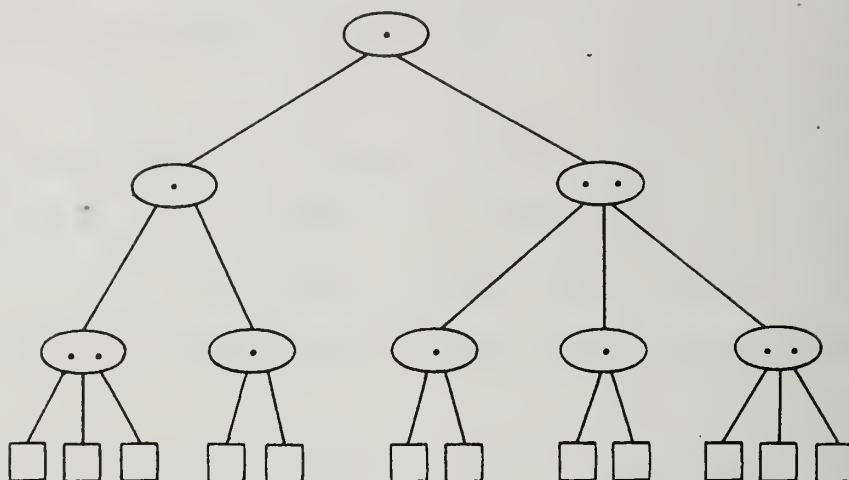


Figure 1 A 3-2 tree with 11 keys



Figure 2 The two types of 3-2 trees of height 1

shall derive bounds on $\bar{n}(N)$ and on the corresponding quantity for B-trees (see Section 3). A systematic procedure for deriving improved bounds is discussed, but the computation involved appears to be prohibitive. Main results are contained in Theorems 2.6, 2.13, and 3.1.

2. Number of Nodes in 3-2 Trees

Let T be any 3-2 tree. We shall use $n(T)$ to denote the number of internal nodes of T . Let $f_N(T)$ be the probability that T will result after N random insertions are made to the empty tree. Obviously $f_N(T)$ is zero unless T contains exactly N keys. In terms of $f_N(T)$ and $n(T)$, the average number of nodes $\bar{n}(N)$ defined in Section 1 can be expressed as follows:

$$\bar{n}(N) = \sum_T n(T) f_N(T) \quad (1)$$

To derive bounds on $\bar{n}(N)$, we observe that most of the internal nodes of any 3-2 tree appear on the lowest few levels. Therefore, a good estimate of $\bar{n}(N)$ can be obtained by analyzing the number of internal nodes in those levels of a random 3-2 tree. We shall carry out the analysis for the lowest level first in the next subsection, and then take the second lowest level into account in Section 2.2.

2.1 First Order Analysis

As shown in Figure 2, there are two types of 3-2 trees of height 1. The type 1 3-2 tree contains 1 keys and the type 2 3-2 tree contains 2 keys. An arbitrary 3-2 tree T is said to be of class $(1; x_1, x_2)$ if among the lowest height 1 subtrees of T , x_1 of them are of type 1 and x_2 are of type 2. The 3-2 tree shown in Figure 1 is of class $(1; 3, 2)$. Let T be an N -key 3-2 tree of class $(1; x_1, x_2)$, the following lemmas are easy to obtain:

Lemma 2.1: $2x_1 + 3x_2 = N+1$

Proof: Both $N+1$ and $2x_1 + 3x_2$ are equal to the number of leaves of T . $\square$

Lemma 2.2: $\frac{3}{2}(x_1 + x_2) - \frac{1}{2} \leq n(T) \leq 2(x_1 + x_2) - 1$

Proof: There are $x_1 + x_2 - 1$ keys contained in the internal nodes above the lowest level. Thus the number of nodes above the lowest level, $n(T) - (x_1 + x_2)$, satisfies $\frac{1}{2}(x_1 + x_2 - 1) \leq n(T) - (x_1 + x_2) \leq x_1 + x_2 - 1$.

Lemma 2.2 follows. $\square$

Definition 2.3: Let $\mathcal{F}(x_1, x_2)$ be the set of 3-2 trees of class $(1; x_1, x_2)$.

Define

$$P_N(x_1, x_2) \stackrel{\text{def}}{=} \sum_{T \in \mathcal{F}(x_1, x_2)} f_N(T)$$

Definition 2.4: $A_i(N) \stackrel{\text{def}}{=} \sum_{x_1, x_2} x_i \cdot P_N(x_1, x_2) \quad i = 1, 2.$

$P_N(x_1, x_2)$ is obviously the probability for a random N -key 3-2 tree to be of type $(1; x_1, x_2)$, and $A_i(N)$ is the average value of x_i for random N -key 3-2 trees.

Lemma 2.5: $\frac{3}{2}(A_1(N) + A_2(N)) - \frac{1}{2} \leq \bar{n}(N) \leq 2(A_1(N) + A_2(N)) - 1.$

Proof: This follows from Lemma 2.2 and the definitions of $\bar{n}(N)$, $A_1(N)$, $A_2(N)$. $\square$

Lemma 2.6: $A_1(N) = \frac{2}{7}(N+1)$, $A_2(N) = \frac{1}{7}(N+1)$ for $N \geq 6$.

Proof: Let T be an $(N-1)$ -key 3-2 tree of class $(1; x_1, x_2)$. By making a random insertion to T , we will obtain a 3-2 tree either of class $(1; x_1-1, x_2+1)$ or of class $(1; x_1+2, x_2-1)$. The former situation happens, with probability $2x_1/N$, when the new key is inserted into a subtree of type 1. Thus we have

$$\begin{aligned}
A_1(N) &= \sum_{x_1, x_2} P_{N-1}(x_1, x_2) \left[\frac{2x_1}{N}(x_1-1) + \left(1 - \frac{2x_1}{N}\right)(x_1+2) \right] \\
&= \sum_{x_1, x_2} P_{N-1}(x_1, x_2) \left(x_1 + \frac{-6x_1}{N} + 2 \right) \\
&= \left(1 - \frac{6}{N}\right) A_1(N-1) + 2
\end{aligned} \tag{2}$$

With initial condition $A_1(1) = 1$, it is easy to show from (2) that

$$A_1(N) = \frac{2}{7}(N+1) \quad \text{for } N \geq 6 \tag{3}$$

Lemma 2.1 implies $2A_1(N) + 3A_2(N) = N+1$. This and (3) give

$$A_2(N) = \frac{1}{7}(N+1) \quad \text{for } N \geq 6 \quad \square$$

Lemma 2.5 and 2.6 lead immediately to the following theorem:

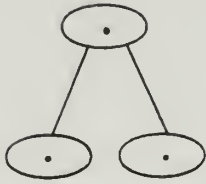
Theorem 2.6: $\frac{9}{14}N + \frac{1}{7} \leq \bar{n}(N) \leq \frac{6}{7}N - \frac{1}{7} \quad \text{for } N \geq 6.$

Corollary: $0.64N \leq \bar{n}(N) \leq 0.86N \quad \text{for } N \geq 6.$

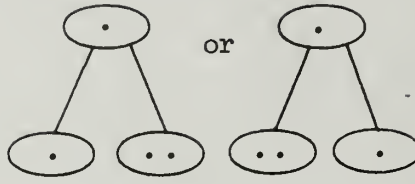
The above bounds should be compared with the obvious bounds $\frac{N}{2} \leq \bar{n}(N) \leq N$, which can be regarded as the zeroth order approximation of $\bar{n}(N)$.

2.2 Second Order Analysis

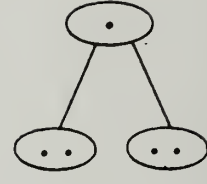
Better bounds for $\bar{n}(N)$ can be derived by considering the internal nodes on the lowest 2 levels of 3-2 trees. Let us divide all 3-2 trees of height 2 into 9 types as shown in Figure 3. For any 3-2 tree T with no fewer than 3 keys, we can classify T by its height 2 subtrees. We shall say that T is of class $(2; x_1, x_2, \dots, x_9)$ if there are x_i height 2 subtrees of type i for each i (Figure 4). Let T be an N -key 3-2 tree of class $(2; x_1, x_2, \dots, x_9)$. The following two lemmas are easy to prove.



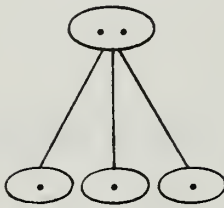
Type 1



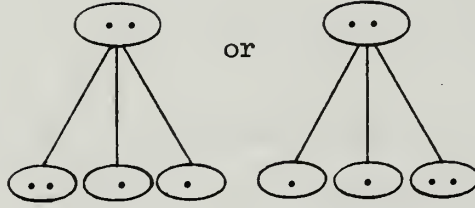
Type 2



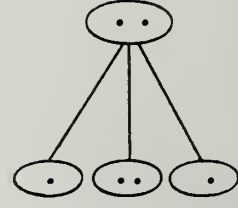
Type 3



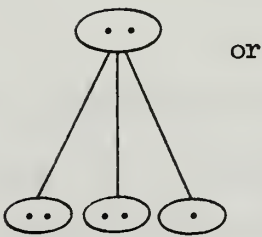
Type 4



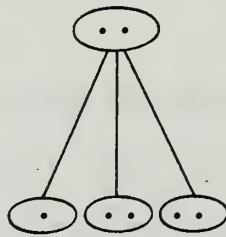
Type 5



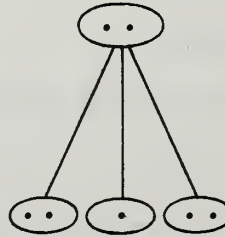
Type 6



Type 7



Type 8



Type 9

Figure 3 There are 9 types of 3-2 trees of height 2 (leaves not shown).

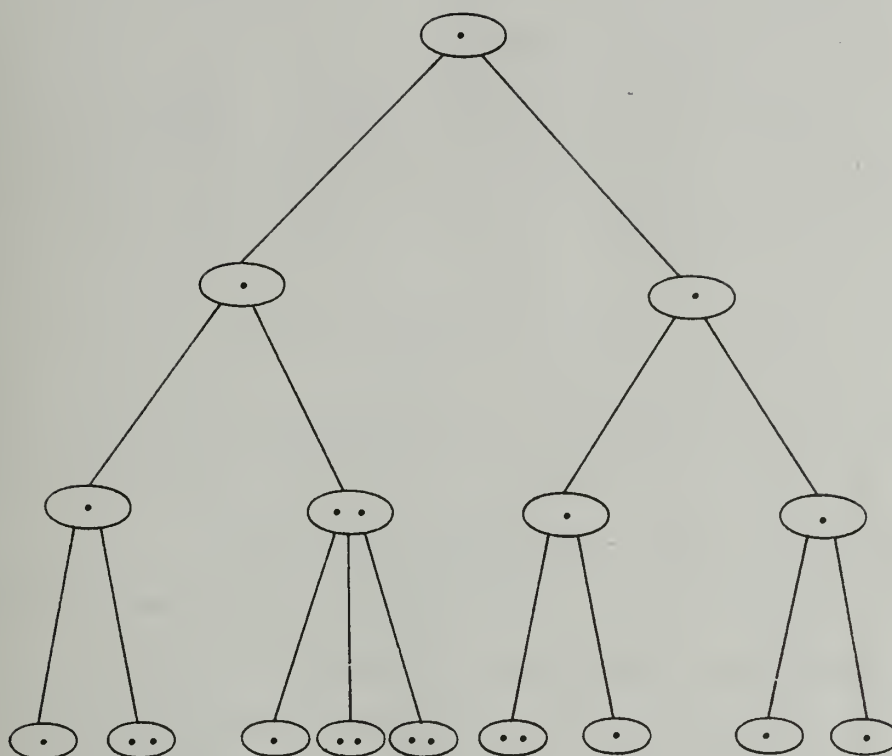


Figure 4 A 3-2 tree of class $(2;1,2,0,0,0,0,1,0,0)$

Lemma 2.7: $4x_1 + 5x_2 + 6x_3 + 6x_4 + 7x_5 + 7x_6 + 8x_7 + 8x_8 + 9x_9 = N+1.$

Proof: Similar to the proof of Lemma 2.1. □

Lemma 2.8: $\frac{7}{2} \sum_{i=1}^3 x_i + \frac{9}{2} \sum_{i=4}^9 x_i - \frac{1}{2} \leq n(T) \leq 4 \sum_{i=1}^3 x_i + 5 \sum_{i=4}^9 x_i - 1.$

Proof: Similar to the proof of Lemma 2.2. □

In analogy with the notation $P_N(x_1, x_2)$ defined in Section 2.1, we use $P_N(2; x_1, x_2, \dots, x_9)$ to denote the probability for an N -key random 3-2 tree to be of class $(2; x_1, x_2, \dots, x_9)$. For each $i (1 \leq i \leq 9)$, define

$$A_i(N) = \sum_{x_1, \dots, x_9} x_i P_N(2; x_1, \dots, x_9).$$

Lemma 2.9: $\frac{7}{2} \sum_{i=1}^3 A_i(N) + \frac{9}{2} \sum_{i=4}^9 A_i(N) - \frac{1}{2} \leq \bar{n}(N) \leq 4 \sum_{i=1}^3 A_i(N) + 5 \sum_{i=4}^9 A_i(N) - 1.$

Proof: Use Lemma 2.8 and definitions of $A_i(N)$, $\bar{n}(N)$. □

We shall study the values of the $A_i(N)$'s. Once these numbers are known, Lemma 2.9 determines $\bar{n}(N)$ to within 15%.

Consider any $(N-1)$ -key 3-2 tree T of class $(2; x_1, x_2, \dots, x_9)$. By examining the insertion process, it can be seen that there are 17 classes of trees that T might become upon the random insertion of a key. These 17 possible classes together with their probabilities of occurrence are tabulated in Table 1. Recurrence relations for the $A_i(N)$'s can be obtained from Table 1 as in Section 2.1. For example, it is easy to show that

$$\begin{aligned} A_1(N) &= \sum_{x_i} P_{N-1}(2; x_1, \dots, x_9) \left[x_1 + \frac{4x_1}{N}(-1) + \frac{3x_5}{N} \cdot (2) + \frac{3x_6}{N} \cdot (2) \right. \\ &\quad \left. + \frac{6x_7}{N} \cdot (1) + \frac{6x_8}{N} \cdot (1) + \frac{6x_9}{N} \cdot (1) \right] \\ &= A_1(N-1) + \frac{1}{N}(-4A_1(N-1) + 6A_5(N-1) + 6A_6(N-1) + 6A_7(N-1) + 6A_8(N-1) \\ &\quad + 6A_9(N-1)). \end{aligned}$$

x_1'	x_2'	x_3'	x_4'	x_5'	x_6'	x_7'	x_8'	x_9'	Probability
x_1-1	x_2+1	x_3	x_4	x_5	x_6	x_7	x_8	x_9	$4x_1/N$
x_1	x_2-1	x_3+1	x_4	x_5	x_6	x_7	x_8	x_9	$2x_2/N$
x_1	x_2-1	x_3	x_4+1	x_5	x_6	x_7	x_8	x_9	$3x_2/N$
x_1	x_2	x_3-1	x_4	x_5+1	x_6	x_7	x_8	x_9	$6x_3/N$
x_1	x_2	x_3	x_4-1	x_5+1	x_6	x_7	x_8	x_9	$4x_4/N$
x_1	x_2	x_3	x_4-1	x_5	x_6+1	x_7	x_8	x_9	$2x_4/N$
x_1	x_2	x_3	x_4	x_5-1	x_6	x_7+1	x_8	x_9	$2x_5/N$
x_1	x_2	x_3	x_4	x_5-1	x_6	x_7	x_8+1	x_9	$2x_5/N$
x_1+2	x_2	x_3	x_4	x_5-1	x_6	x_7	x_8	x_9	$3x_5/N$
x_1	x_2	x_3	x_4	x_5	x_6-1	x_7+1	x_8	x_9	$4x_6/N$
x_1+2	x_2	x_3	x_4	x_5	x_6-1	x_7	x_8	x_9	$3x_6/N$
x_1	x_2	x_3	x_4	x_5	x_6	x_7-1	x_8	x_9+1	$2x_7/N$
x_1+1	x_2+1	x_3	x_4	x_5	x_6	x_7-1	x_8	x_9	$6x_7/N$
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8-1	x_9+1	$2x_8/N$
x_1+1	x_2+1	x_3	x_4	x_5	x_6	x_7	x_8-1	x_9	$6x_8/N$
x_1+1	x_2	x_3+1	x_4	x_5	x_6	x_7	x_8	x_9-1	$6x_9/N$
x_1	x_2+2	x_3	x_4	x_5	x_6	x_7	x_8	x_9-1	$3x_9/N$

Table 1 Transition under a random insertion:
 a tree of class $(2; x_1, x_2, \dots, x_9)$
 becomes a tree of class $(2; x_1', x_2', \dots, x_9')$.
 Each row gives the values of $x_1', x_2', \dots, x_9'$
 for a possible resulting class with its
 probability of occurrence in the last
 column.

Similar formulas for $A_2(N), \dots, A_9(N)$ can also be derived. These relations can be compactly written in the following form: Let $A(N)$ be the 9-component column vector $(A_i(N))$, then

$$A(N) = (I + \frac{1}{N}D)A(N-1) \quad (4)$$

where I is the 9×9 identity matrix and D is given by

$$D = \begin{pmatrix} -4 & 0 & 0 & 0 & 6 & 6 & 6 & 6 & 6 \\ 4 & -5 & 0 & 0 & 0 & 0 & 6 & 6 & 6 \\ 0 & 2 & -6 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 3 & 0 & -6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 4 & -7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & -7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 4 & -8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & -8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & -9 \end{pmatrix} \quad (5)$$

To solve $A(N)$ from (4), we define a 9-component column vector $a(N) = (a_i(N))$ by

$$a_i(N) = A_i(N)/(N+1) \quad (6)$$

In terms of $a(N)$, (4) can be written as

$$a(N) = (I + \frac{1}{N+1}(D-I))a(N-1) \quad (7)$$

To solve (7), the following two lemmas will be useful. Since recurrence relation of the form (7) has been studied in some other context [5], we shall omit the proofs here.

Lemma 2.10: Let G be a $p \times p$ real matrix with simple eigenvalues

$\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_{p-1}$ where $\lambda_0 = 0$ and $\text{Re}\lambda_{p-1} \leq \text{Re}\lambda_{p-2} \leq \dots \leq \text{Re}\lambda_1 < 0$.

If $v(1), v(2), \dots, v(N), \dots$ is a sequence of p -component vectors satisfying

$v(j) = (1 + \frac{1}{j+1}G)v(j-1)$, then there exists a vector u such that

- (i) $Gu = 0$
(ii) $|(v(N))_i - u_i| < CN^{\operatorname{Re} \lambda_1}$ for some constant C and all i, N .

where $(v(N))_i, u_i$ denote the i^{th} component of $v(N)$ and u respectively.

Corollary: $v(N) \rightarrow u$ as $N \rightarrow \infty$.

Lemma 2.11: Let k be a positive integer, then the polynomial

$$g(\lambda) = \prod_{j=0}^{\ell} (\lambda+k+j) - \prod_{j=0}^{\ell} (k+j) \text{ has only simple roots. Furthermore,}$$

the real parts of all the roots except the root $\lambda = 0$ are negative.

We now turn to the determination of $a(N)$ by (7). An explicit calculation shows that the characteristic polynomial of $D-I$ is $-\lambda(\lambda+7)(\lambda+8)(\lambda+9)((\lambda+9)^5 + 25(\lambda+9)^3 + 210(\lambda+9)^2 + 136(\lambda+9) + 384)$. The roots of the polynomial, which are eigenvalues of $D-I$, are $0, -6.55 \pm 6.25i, -7, -8, -9, -9.23 \pm 1.37i, -13.44$. Thus, $D-I$ satisfies the conditions on G in Lemma 2.10. Therefore, there exists a vector $u = (u_i)$ such that

$$|a_i(N) - u_i| < C_0 N^{-6.55} \quad \text{for some constant } C_0 \quad (8)$$

where u satisfies

$$(D-I)u = 0 \quad (9)$$

In terms of the u_i 's, we can express Lemma 2.9 as follows:

Lemma 2.12:

$$(N+1)\left(\frac{7}{2} \sum_{i=1}^3 u_i + \frac{9}{2} \sum_{i=4}^9 u_i\right) - \frac{1}{2} - CN^{-5.55} \leq \bar{n}(N) \leq (N+1)\left(4 \sum_{i=1}^3 u_i + 5 \sum_{i=4}^9 u_i\right) - 1 + CN^{-5.55}$$

for some constant C .

Proof: From equations (6) and (8), we obtain

$$|A_i(N) - (N+1)u_i| < \frac{C}{9} N^{-5.55} \quad \text{for some constant } C. \quad (10)$$

The lemma follows immediately from (10) and Lemma 2.9. $\square$

Now, to find the values of the u_i 's, we observe that equation (9) determines u up to a constant factor. The normalization constant can be determined as follows: Lemma 2.7 and equation (6) lead to the equation

$$4a_1(N) + 5a_2(N) + 6a_3(N) + 6a_4(N) + 7a_5(N) + 7a_6(N) + 8a_7(N) + 8a_8(N) + 9a_9(N) = 1 \quad (11)$$

Since $a_i(N) \rightarrow u_i$ as $N \rightarrow \infty$, (11) implies that

$$4u_1 + 5u_2 + 6u_3 + 6u_4 + 7u_5 + 7u_6 + 8u_7 + 8u_8 + 9u_9 = 1, \quad (12)$$

which is the equation we need in order to determine the normalization constant. Therefore, solving equations (9) and (12), we obtain:

$$\begin{aligned} u_1 &= 414/7991 = 0.052 \\ u_2 &= 396/7991 = 0.050 \\ u_3 &= 912/55937 = 0.016 \\ u_4 &= 1188/55937 = 0.021 \\ u_5 &= 1278/55937 = 0.023 \\ u_6 &= 297/55937 = 0.005 \\ u_7 &= 416/55937 = 0.007 \\ u_8 &= 284/55937 = 0.005 \\ u_9 &= 20/7991 = 0.003 \end{aligned} \quad (13)$$

Substituting the values of the u_i 's into the inequality in Lemma 2.12, we obtain the main result of this section.

Theorem 2.13: $0.70N + 0.2 - CN^{-5.55} \leq \bar{n}(N) \leq 0.79N - 0.2 + CN^{-5.55}$ for some constant C .

The technique used in this subsection can be used to compute higher moments of the number of nodes $n(T)$. For example, we can set up a system of recurrence relations of the form of equation (4) for the quantities $A_{ij}(N)$ where

$$A_{ij}(N) = x_1, \dots, x_9 \sum P_N(2; x_1, \dots, x_9) x_i x_j \quad i, j = 1, 2, \dots, 9.$$

Determination of the $A_{ij}(N)$'s will then lead to (by Lemma 2.8) bounds on the average value of $n(T)^2$ for N -key random 3-2 trees.

2.3 Higher Order Analysis

The methods used in the previous two subsections obviously can be generalized to obtain better approximation of $\bar{n}(N)$. By computing the average number of nodes in the lowest k -levels, we can determine $\bar{n}(N)$ to an accuracy of $1/2(2^k - 1) \times 100\%$. This is so because at most $1/2^k$ of the keys are in nodes above the lowest k levels.

This general procedure, however, is not very useful in practice. If $F(k)$ is the number of different types of trees of height k , then solution of this problem involves the manipulation of an $F(k) \times F(k)$ matrix. It is not difficult to show that $F(k) = \frac{1}{2}F(k-1)(F(k-1)+1)^2$. We have analyzed the cases $k = 1, 2$, where $F(1) = 2$ and $F(2) = 9$. However, $F(3) = 450$ is already such a large number that carrying out the computation appears to be very difficult.

3. An Analysis of B-trees

3.1 Introduction

A natural extension of 3-2 trees is the idea of "B-trees" [2] [4]. A B-tree of order m is a tree in which the number of keys contained

in any internal node other than the root is no greater than $m-1$ and no less than $\lceil m/2 \rceil - 1$. 3-2 trees are just B-trees of order 3. To add a key to a node, we insert the new key into the other keys and check if the node now contains more than $m-1$ keys. If the answer is no, the insertion has been completed. Otherwise, we split the node into 2 nodes, one of which contains the smallest $\lceil m/2 \rceil - 1$ keys and the other the $m - \lceil m/2 \rceil$ largest keys, the one remaining key is then inserted into the parent node. Random B-trees are defined in exactly the same way as random 3-2 trees are defined.

We shall study $\bar{n}_m(N)$, the average number of nodes in the B-trees of order m resulting from N random insertions. An obvious bound was given in [4]:

$$\frac{1}{m-1} \leq \frac{\bar{n}_m(N)}{N} \leq \frac{1}{\lceil m/2 \rceil - 1} + \frac{1}{N} \quad (14)$$

In this section we shall consider the nodes at the lowest level, and do an analysis similar to the first order analysis done in Section 2.1. As we shall see, this analysis yields better results than the corresponding analysis for 3-2 trees. This is so because a greater proportion of keys in a B-tree are stored in the lowest internal nodes as m becomes larger.

Define the following functions:

$$H(N) \stackrel{\text{def}}{=} \sum_{k=1}^N \frac{1}{k} \quad \text{for } N \geq 1$$

$$r(m) \begin{cases} \stackrel{\text{def}}{=} \frac{1}{m+1} (H(m) - H(m/2))^{-1} & \text{if } m = \text{even} \\ \stackrel{\text{def}}{=} \frac{1}{m+1} (H(m+1) - H((m+1)/2))^{-1} & \text{if } m = \text{odd} \end{cases}$$

It is well known [6] that $H(m) \sim \ln m + 0.58 + \frac{1}{2m} + \dots$. A simple computation shows that $r(m) = \frac{1}{m} \frac{1}{\ln 2} + O(\frac{1}{m^2})$. Our new bounds on $\bar{n}_m(N)$ are given below:

Theorem 3.1: For any $\epsilon > 0$ and fixed m ,

$$(1 + \frac{1}{m-1})r(m) - \epsilon \leq \frac{\bar{n}_m(N)}{N} \leq (1 + \frac{1}{\lceil m/2 \rceil - 1})r(m) + \epsilon$$

when N is sufficiently large.

Corollary: For any $\epsilon > 0$ and fixed m , $|\frac{\bar{n}_m(N)}{N} - \frac{1}{m} \frac{1}{\ln 2}| < \frac{C}{m^2} + \epsilon$ for all sufficiently large N , where C is a constant independent of m and N .

The corollary follows from Theorem 3.1 and the approximation of $r(m)$ given earlier. If all the nodes in a B-tree of order m contain $m-1$ keys, there would be $N/(m-1)$ nodes. The ratio $N/(m-1)\bar{n}_m(N)$ can therefore be viewed as storage utilization [4]. Our corollary to Theorem 3.1 shows that, as N becomes large, the storage utilization is essentially $\ln 2 \sim 0.69$ for fixed large m (cf. equation (14)).

3.2 Proof of Theorem 3.1

We will first introduce some notations. Note that there are $m - \lceil m/2 \rceil + 1$ types of B-trees of order m and height 1. As shown in Figure 5, a type i B-tree contains i keys in its node for $i = \lceil m/2 \rceil - 1, \lceil m/2 \rceil, \dots, m-1$. A B-tree of order m is said to be of class $(y_{\lceil m/2 \rceil - 1}, y_{\lceil m/2 \rceil}, \dots, y_{m-1})$ if at the lowest level there are y_i subtrees of type i for each i . Let $P_N(y_{\lceil m/2 \rceil - 1}, y_{\lceil m/2 \rceil}, \dots, y_{m-1})$ be the probability for a random N -key B-tree of order m to be of class $(y_{\lceil m/2 \rceil - 1}, \dots, y_{m-1})$.

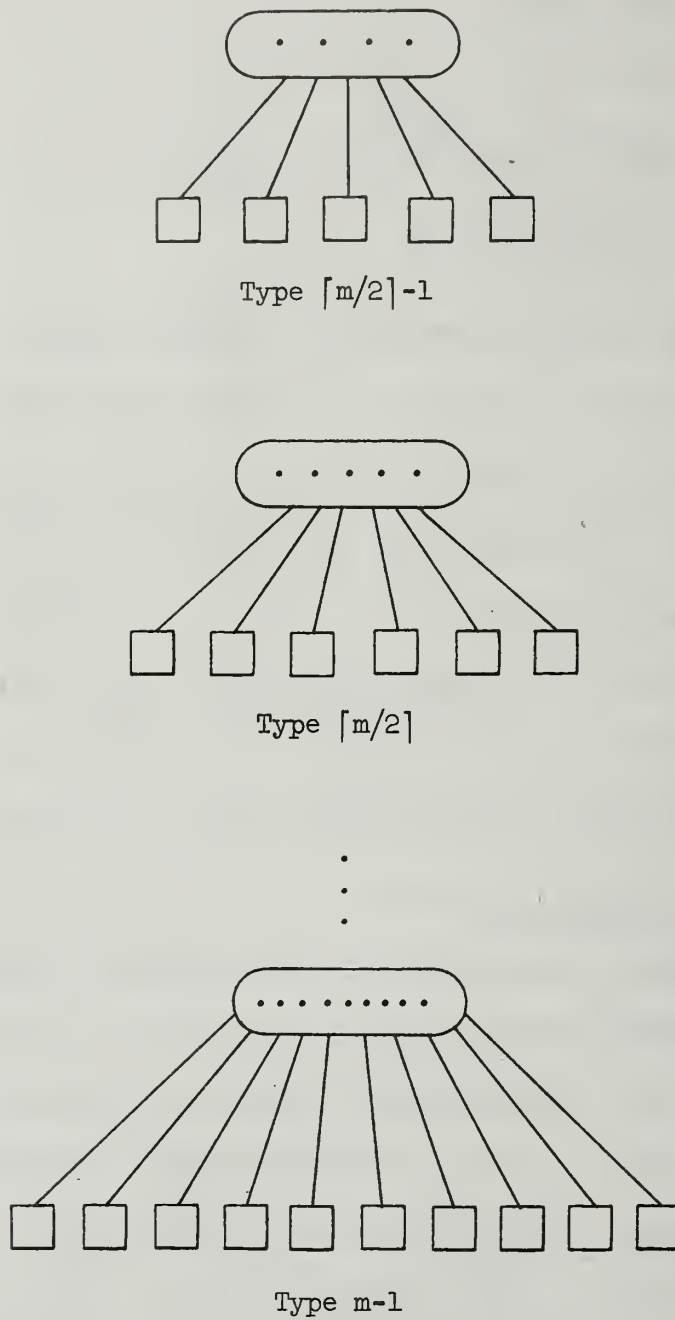


Figure 5 The height 1 B-trees of order m consist of $m - \lfloor m/2 \rfloor + 1$ types.
(Shown for $m = 10$)

Definition 3.2: $A_i(N) \stackrel{\text{def}}{=} \sum_{j=1}^m y_j P_N(y_{\lceil m/2 \rceil - 1}, \dots, y_{m-1})$

$$i = \lceil m/2 \rceil - 1, \dots, m-1.$$

For brevity, we have suppressed the dependence of $A_i(N)$ and P_N on m in our notations.

$$\begin{aligned} \text{Lemma 3.3: } \left(1 + \frac{1}{m-1}\right) \sum_{i=\lceil m/2 \rceil - 1}^{m-1} A_i(N) - \frac{1}{m-1} &\leq \bar{n}_m(N) \\ &\leq \left(1 + \frac{1}{\lceil m/2 \rceil - 1}\right) \sum_{i=\lceil m/2 \rceil - 1}^{m-1} A_i(N) - \frac{1}{\lceil m/2 \rceil - 1} + 1 \end{aligned}$$

Proof: Similar to the proof of Lemma 2.1. The term $+1$ appearing on the right-hand side of the equation arises from the fact that the root may contain less than $\lceil m/2 \rceil - 1$ keys. $\square$

The major effort to prove Theorem 3.1 is contained in the next Lemma.

Lemma 3.4: Let $g(N) \stackrel{\text{def}}{=} \sum_{i=\lceil m/2 \rceil - 1}^{m-1} A_i(N)/(N+1)$. Then for any $\epsilon > 0$,

$$|g(N) - r(m)| < \epsilon \quad \text{for all sufficiently large } N.$$

Proof: We shall assume $m = 2p$ to be an even number. The proof for odd m is similar.

Let T be an $(N-1)$ -key B-tree of order $2p$. After a random insertion, T may become a B-tree of class $(y_{p-1}, \dots, y_{i-1}^{-1}, y_i^{+1}, \dots, y_{2p-1})$ with probability iy_{i-1}/N (for each $i = p, \dots, 2p-1$) or it may become a B-tree of class $(y_{p-1}^{+1}, y_p^{+1}, y_{p+1}, \dots, y_{2p-1}^{-1})$ with probability $2py_{2p-1}/N$. It follows that

$$A_{p-1}(N) = A_{p-1}(N-1) + \frac{1}{N}(-pA_{p-1}(N-1) + 2pA_{2p-1}(N-1))$$

$$A_p(N) = A_p(N-1) + \frac{1}{N}(pA_{p-1}(N-1) - (p+1)A_p(N-1) + 2pA_{2p-1}(N-1))$$

and

$$A_j(N) = A_j(N-1) + \frac{1}{N}(jA_{j-1}(N-1) - (j+1)A_j(N-1))$$

(15)

for $p+1 \leq j \leq 2p-1$

Denoting by $A(N)$ the $(p+1)$ -component vector $(A_j(N))$, (15) can be written in matrix notation as

$$A(N) = (I + \frac{1}{N}B)A(N-1) \quad (16)$$

where I is the $(p+1) \times (p+1)$ identity matrix, and B is defined by

$$B = \begin{pmatrix} -p & & & & & 2p \\ & p & -(p+1) & & & 2p \\ & & p+1 & -(p+2) & & 0 \\ & & & p+2 & -(p+3) & 0 \\ & & & & \ddots & \vdots \\ & & & & & 2p-1 & -2p \end{pmatrix} \quad (17)$$

To solve (16) for $A(N)$, we define a $(p+1)$ -component vector $a(N) = (a_i(N))$ by

$$a_i(N) = A_i(N)/(N+1) \quad (18)$$

Equations (16) and (18) lead to the following recurrence relation

$$a(N) = [I + \frac{1}{N+1}(B-I)]a(N-1) \quad (19)$$

The characteristic polynomial $q(\lambda)$ of $B-I$ is computed to be

$$q(\lambda) = (-1)^{p+1}(\lambda+2p+1) \left[\sum_{j=1}^p (\lambda+p+j) - \sum_{j=1}^p (p+j) \right]$$

From Lemma 2.11, it is easy to see that the roots of $q(\lambda) = 0$ satisfy the following conditions:

- (i) All roots are simple roots.
- (ii) $\lambda = 0$ is a root.
- (iii) The real parts of all roots except $\lambda = 0$ are negative.

Therefore, according to Lemma 2.10, there exists a vector

$u = (u_{p-1}, u_p, \dots, u_{2p-1})^T$ such that

$$(i) \quad (B-I) u = 0 \quad (21)$$

$$(ii) \quad |a_i(N) - u_i| < C_m N^{-\epsilon_m} \quad \text{for } p-1 \leq i \leq 2p-1 \quad (22)$$

where C_m, ϵ_m are positive constants.

(22) implies

$$\left| \sum_{i=p-1}^{2p-1} A_i(N)/(N+1) - \sum_{i=p-1}^{2p-1} u_i \right| \leq C'_m N^{-\epsilon_m} \quad (23)$$

Now, to determine the u_i 's, we note that the following equation can be proved easily (cf. the derivation of Equation (12)):

$$\sum_{i=p-1}^{2p-1} (i+1) u_i = 1 \quad (24)$$

Solving (21) and (24), we obtain

$$u_{p-1} = \frac{1}{p+1} \frac{1}{2p+1} [H(2p) - H(p)]^{-1} \quad (25)$$

$$u_i = \frac{1}{i+1} \frac{1}{i+2} [H(2p) - H(p)]^{-1} \quad p \leq i \leq 2p-1$$

$$\text{Therefore, } \sum_{i=p-1}^{2p-1} u_i = \frac{1}{2p+1} (H(2p) - H(p))^{-1} = r(m). \quad (26)$$

Finally, substituting (26) in (23), the lemma is obtained. $\square$

Proof of Theorem 3.1: It is a direct consequence of Lemma 3.3 and 3.4. $\square$

4. Conclusion

We have derived bounds on the average number of nodes in an N-key random B-tree, which essentially is the average number of "splitting" in building the tree. One interesting result is that the asymptotic storage utilization is approximately $\ln 2 \sim 69\%$ for B-trees of high orders. This seems to agree well with one set of experimental data ($m = 121$, $N = 5000$, storage utilization = 67% , see [4]).

Many problems about 3-2 trees remain to be investigated. What is the average number of splitting on the N^{th} random insertion [3]? How to analyze 3-2 trees when deletions are also present? Some upper bounds can be obtained for the former problem using the present approach, but it appears that very different methods would be required to answer these questions satisfactorily.

References

- [1] Aho, Hopcroft, Ullman, The Design and Analysis of Computer Algorithms, Addison-Wesley, 1974.
- [2] Bayer, R. and E. McCreight, Acta Informatica (1972), pp. 173-189.
- [3] Chvatal, V., D. A. Klarner, and D. E. Knuth, Selected Combinatorial Research Problems, Problem 37, STAN-CS-72-292, 1972.
- [4] Knuth, D. E., The Art of Computer Programming, Vol. 3, Addison-Wesley, 1973, pp. 468-480.
- [5] Knuth, D. E., The Art of Computer Programming, Vol. 3, Addison-Wesley, 1973, pp. 679-680, answer to exercise 10.
- [6] Knuth, D. E., The Art of Computer Programming, Vol. 1, Addison-Wesley, 1968, p. 74.

BIBLIOGRAPHIC DATA SHEET	1. Report No. UIUCDCS-R-74-679	2.	3. Recipient's Accession No.
	Title and Subtitle On Random 3-2 Trees		5. Report Date October 1974
Author(s) Andrew Chi-Chih Yao		6.	
Performing Organization Name and Address Department of Computer Science University of Illinois at Urbana-Champaign Urbana, Illinois 61801		8. Performing Organization Rept. No.	
Sponsoring Organization Name and Address National Science Foundation Washington, D.C.		10. Project/Task/Work Unit No.	
		11. Contract/Grant No. NSF GJ-41538	
		13. Type of Report & Period Covered	
Supplementary Notes		14.	
Abstracts It is shown that $\bar{n}(N)$, the average number of nodes in an N-key random 3-2 tree, satisfies the inequality $0.70N - \epsilon < \bar{n}(N) < 0.79N + \epsilon$. A similar analysis is done for general B-trees. It is shown that storage utilization is essentially $0.69 \approx 69\%$ for B-trees of high orders.			
Key Words and Document Analysis. 17a. Descriptors			
Identifiers/Open-Ended Terms			
COSATI Field/Group			
Availability Statement		19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages
		20. Security Class (This Page) UNCLASSIFIED	22. Price

SEP 5 1975

UNIVERSITY OF ILLINOIS-URBANA

510.64 IL6R no. C002 no.679-684(1974

Report /



3 0112 088401515